逻辑回归

## 一、实验内容：

利用决策树模型完成链接<https://archive.ics.uci.edu/ml/datasets/HIV-1+protease+cleavage>的任务

## 二、实验任务：

一共有四个数据集，分别用以其中一个数据集作为训练集，其他三个数据集作为测试集，测试模型性能。下面为四个数据集的相关信息。

- 746 data: 401 cleaved, 345 non-cleaved

- 1625 data: 374 cleaved, 1251 non-cleaved

- Schilling data: 434 cleaved, 2838 non-cleaved

- Impens data: 149 cleaved, 798 non-cleaved

数据集的属性是8个字母的字符串（每个字母代表一个独立的属性）。'ARNDCQEGHILKMFPSTWYV'一共有20个供选择的值，分别代表20个不同的氨基酸。要预测的是此字符串是否代表HIV-1蛋白酶对其进行切割的肽段（或蛋白质）中的位点（如果是，则为+1，否则为-1）。

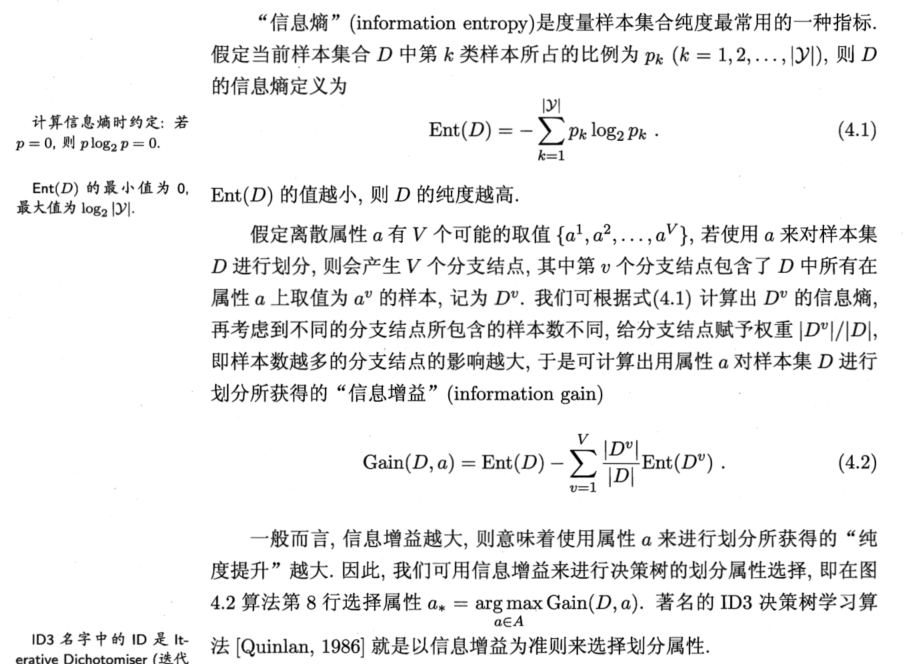
此时，我们的任务可以用下列式子表示：

## 三、实验算法：

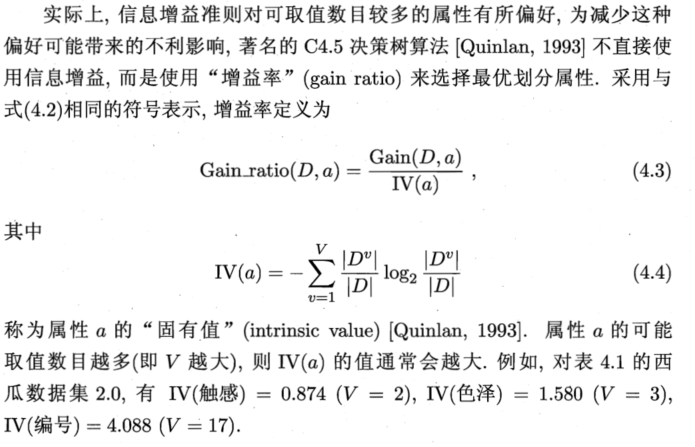
**决策树模型基于树结构进行决策，是一种条件判断，条件规则组合的算法。它根据分层和分割的方式，将特征属性空间划分为一系列简单区域。一棵决策树包含一个根结点、若干个内部结点和若干个叶子结点；叶结点对应于决策结果，其他每个节点对应于一个属性测试。**

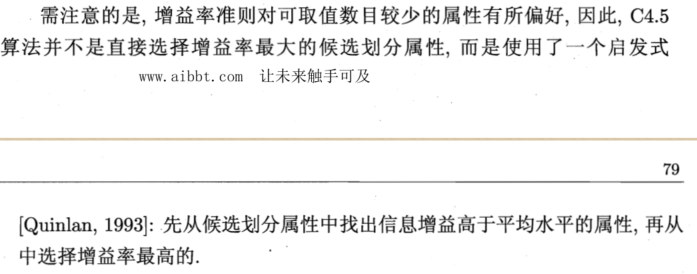
**我们希望选择最优的划分属性使得决策树分支结点所包含的样本尽可能属于同一类别，即结点的纯度越高越好。**

### 1、信息增益

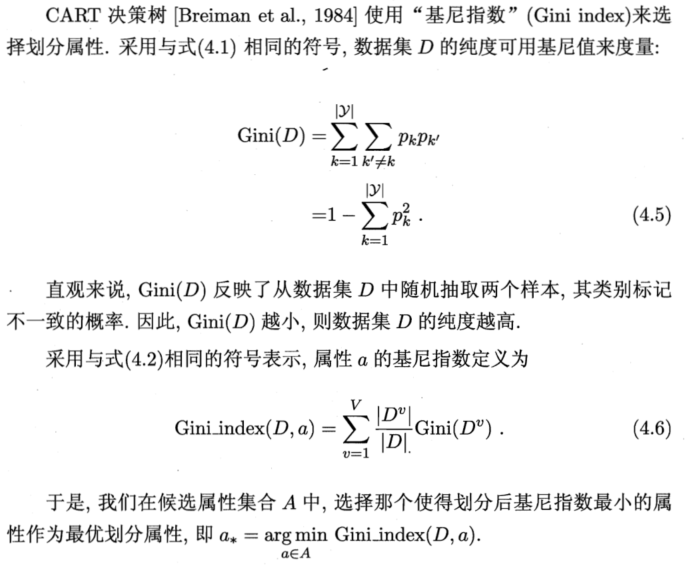


### 2、增益率





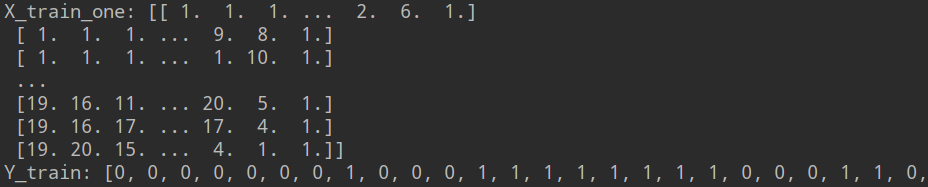
### 3、基尼指数



## 四、实验数据处理：

**一共有4个数据集，按照实验要求，以任意一个数据集作为训练集，以其他数据集作为测试集，来对模型性能进行评估。**

**考虑到每条数据的属性都是字符型，需要将其离散数值化。**'ARNDCQEGHILKMFPSTWYV'将其分别数值化为，1，2，3…，20。



## 五、实验结果：

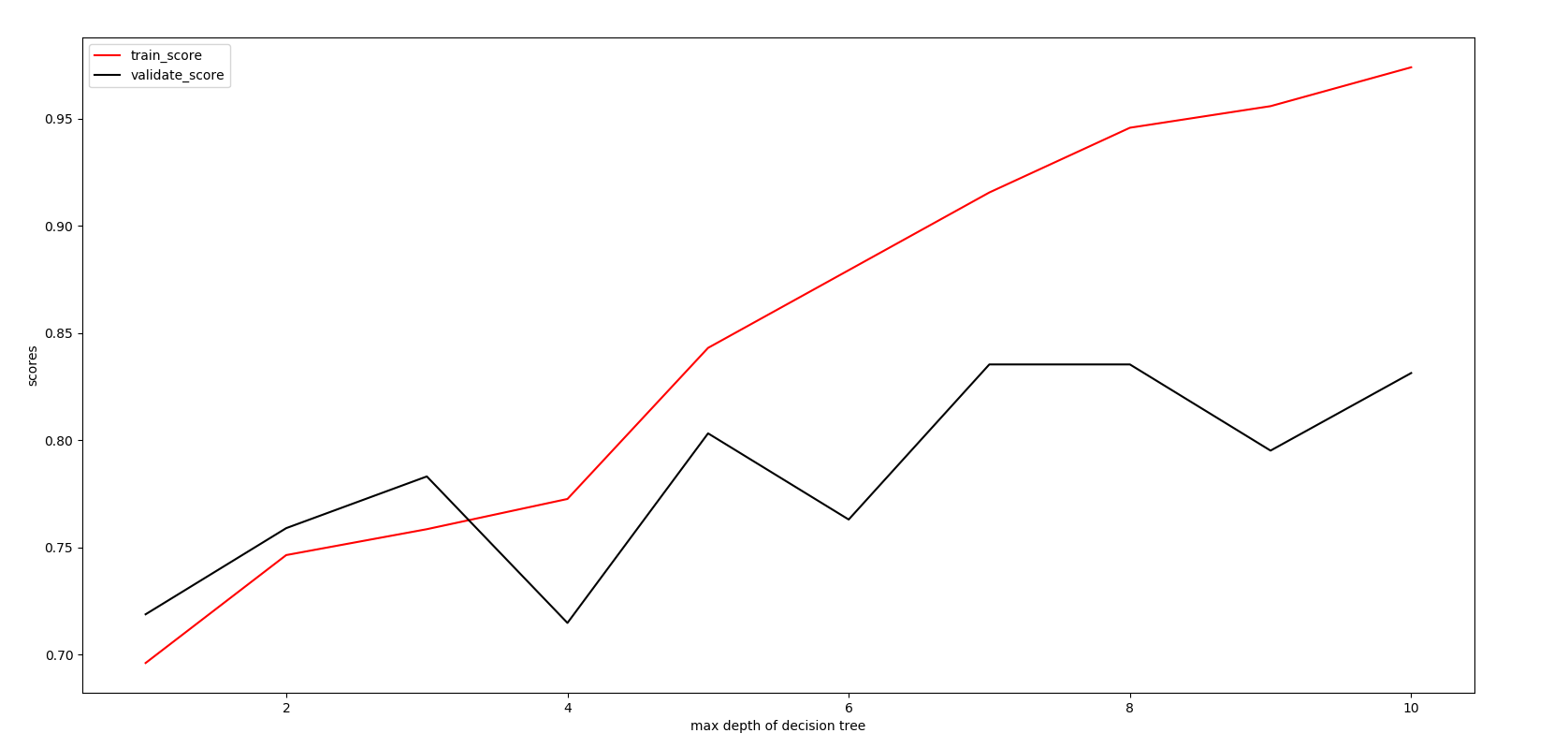
### 1、通过sklearn包实现

Sklearn.tree.DecisionTreeClassifier

1. criterion：选择信息增益entropy还是基尼指数gini
2. max\_path：制定树的最大深度
3. splitter：best最优的分裂策略，random最好的随机切分策略
4. Class\_weight=None:类型权重参数，选择为‘balanced’自动根据y值计算类型权重。权重与输入数据中的类频率成反比。
5. min\_impurity\_decrease: 如果节点的分裂导致不纯度的减少(分裂后样本比分裂前更加纯净)大于或等于min\_impurity\_decrease，则分裂该节点。

#### 确定最优的max\_path

1. 对\_746Data进行划分，选取2/3的数据集来学习模型， 选取1/3的数据集来作为验证集，其他数据集作为测试集。使用参数，criterion=‘entropy’， class\_weight=’balanced’,其他为默认参数，确定最优的max\_path。



根据上图，确定当max\_depth为8的时候最佳，依次为参数，对其他数据集进行测试

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| entropy | 训练集 | 验证集 | 测试集 | | |
| 746Data | 2/3的数据 | 1/3的数据 | 1625Data | impensData | schillingData |
|  | 0.9457 | 0.8353 | 0.8425 | 0.7793 | 0.7940 |

max\_depth: 7

train score: 0.9154929577464789

validate score: 0.8353413654618473

test score 1: 0.7827692307692308

test score 2: 0.7085533262935586

test score 3: 0.7435819070904646

----------------------------

max\_depth: 8

train score: 0.9456740442655935

validate score: 0.8353413654618473

test score 1: 0.8424615384615385

test score 2: 0.7793030623020063

test score 3: 0.7940097799511002

----------------------------

max\_depth: 9

train score: 0.9557344064386318

validate score: 0.7951807228915663

test score 1: 0.8652307692307692

test score 2: 0.7845828933474129

test score 3: 0.8080684596577017

----------------------------

max\_depth: 10

train score: 0.9738430583501007

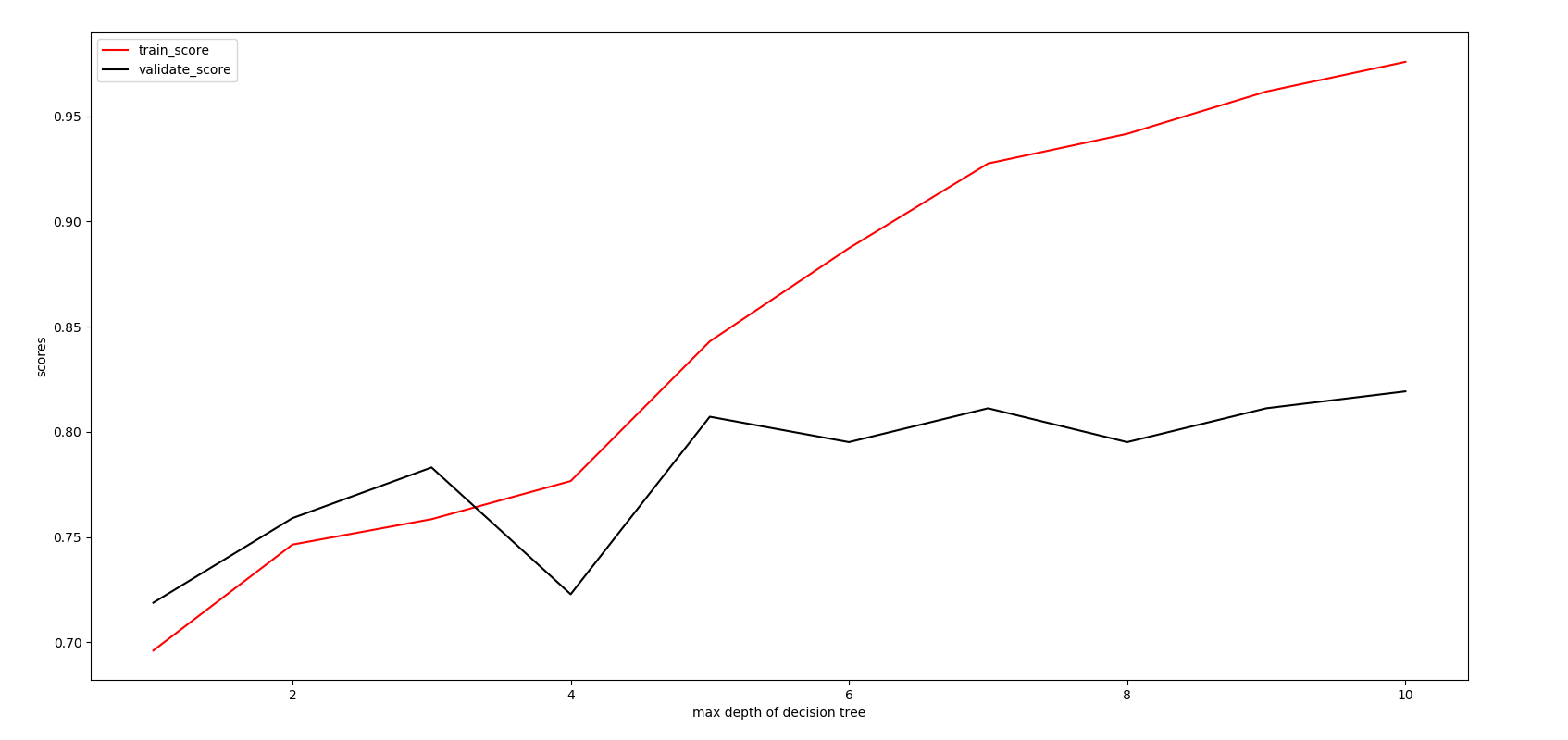
validate score: 0.8313253012048193

test score 1: 0.8486153846153847

test score 2: 0.7740232312565998

test score 3: 0.7897310513447433

1. 对746Data进行划分，选取2/3的数据集来学习模型， 选取1/3的数据集来作为验证集，其他数据集作为测试集。使用参数，criterion=‘gini’， class\_weight=’balanced’,其他为默认参数，确定最优的max\_path。



max\_depth: 7

train score: 0.9275653923541247

validate score: 0.8112449799196787

test score 1: 0.8732307692307693

test score 2: 0.7940865892291447

test score 3: 0.7964547677261614

----------------------------

max\_depth: 8

train score: 0.9416498993963782

validate score: 0.7951807228915663

test score 1: 0.8732307692307693

test score 2: 0.7803590285110876

test score 3: 0.7930929095354523

----------------------------

max\_depth: 9

train score: 0.9617706237424547

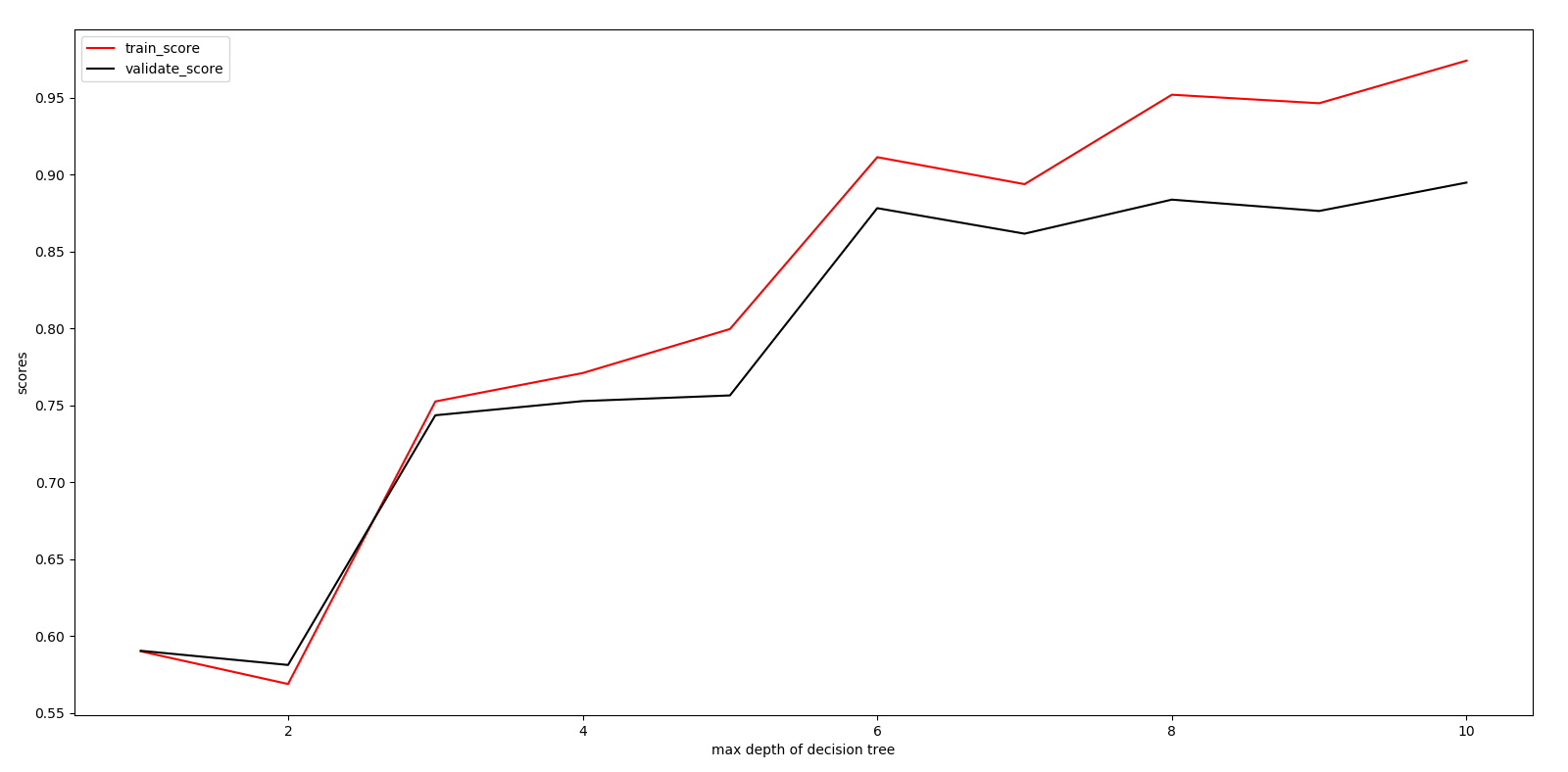
validate score: 0.8112449799196787

test score 1: 0.8363076923076923

test score 2: 0.7550158394931362

test score 3: 0.7735330073349633

1. 对1625Data进行划分，选取2/3的数据集来学习模型， 选取1/3的数据集来作为验证集，其他数据集作为测试集。使用参数，criterion=‘entropy’， class\_weight=’balanced’,其他为默认参数，确定最优的max\_path。



max\_depth: 8

train score: 0.9519852262234534

validate score: 0.8837638376383764

test score 1: 0.8739946380697051

test score 2: 0.8099260823653643

test score 3: 0.8337408312958435

----------------------------

max\_depth: 9

train score: 0.9464450600184672

validate score: 0.8763837638376384

test score 1: 0.8954423592493298

test score 2: 0.8152059134107709

test score 3: 0.8297677261613692

----------------------------

max\_depth: 10

train score: 0.974145891043398

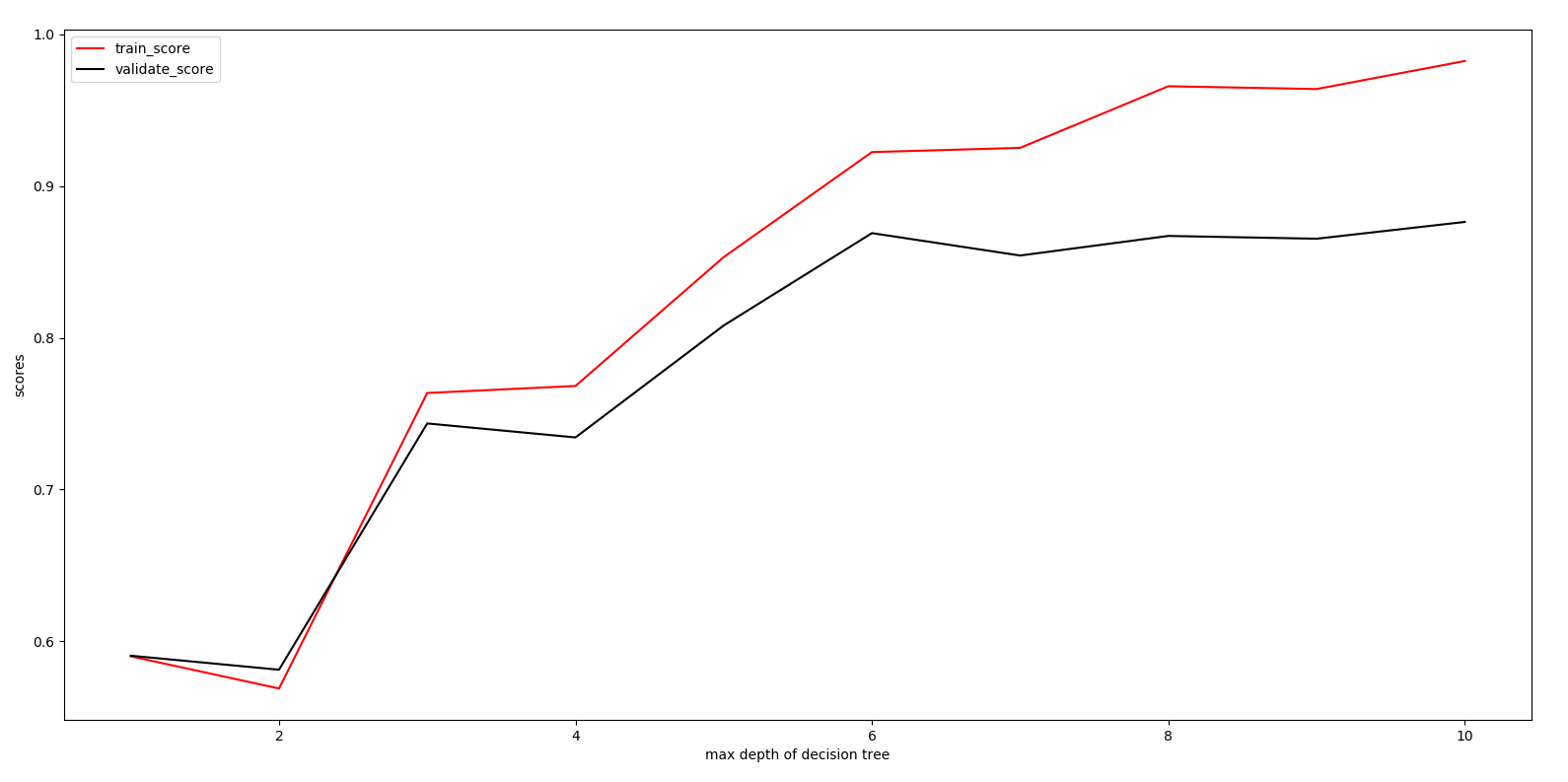
validate score: 0.8948339483394834

test score 1: 0.8954423592493298

test score 2: 0.8204857444561774

test score 3: 0.8447432762836186

1. 对1625Data进行划分，选取2/3的数据集来学习模型， 选取1/3的数据集来作为验证集，其他数据集作为测试集。使用参数，criterion=‘gini’， class\_weight=’balanced’,其他为默认参数，确定最优的max\_path。



max\_depth: 6

train score: 0.9224376731301939

validate score: 0.8690036900369004

test score 1: 0.8646112600536193

test score 2: 0.8035902851108765

test score 3: 0.812958435207824

----------------------------

max\_depth: 7

train score: 0.925207756232687

validate score: 0.8542435424354243

test score 1: 0.8793565683646113

test score 2: 0.791974656810982

test score 3: 0.8056234718826406

----------------------------

max\_depth: 8

train score: 0.9658356417359187

validate score: 0.8671586715867159

test score 1: 0.8806970509383378

test score 2: 0.7961985216473073

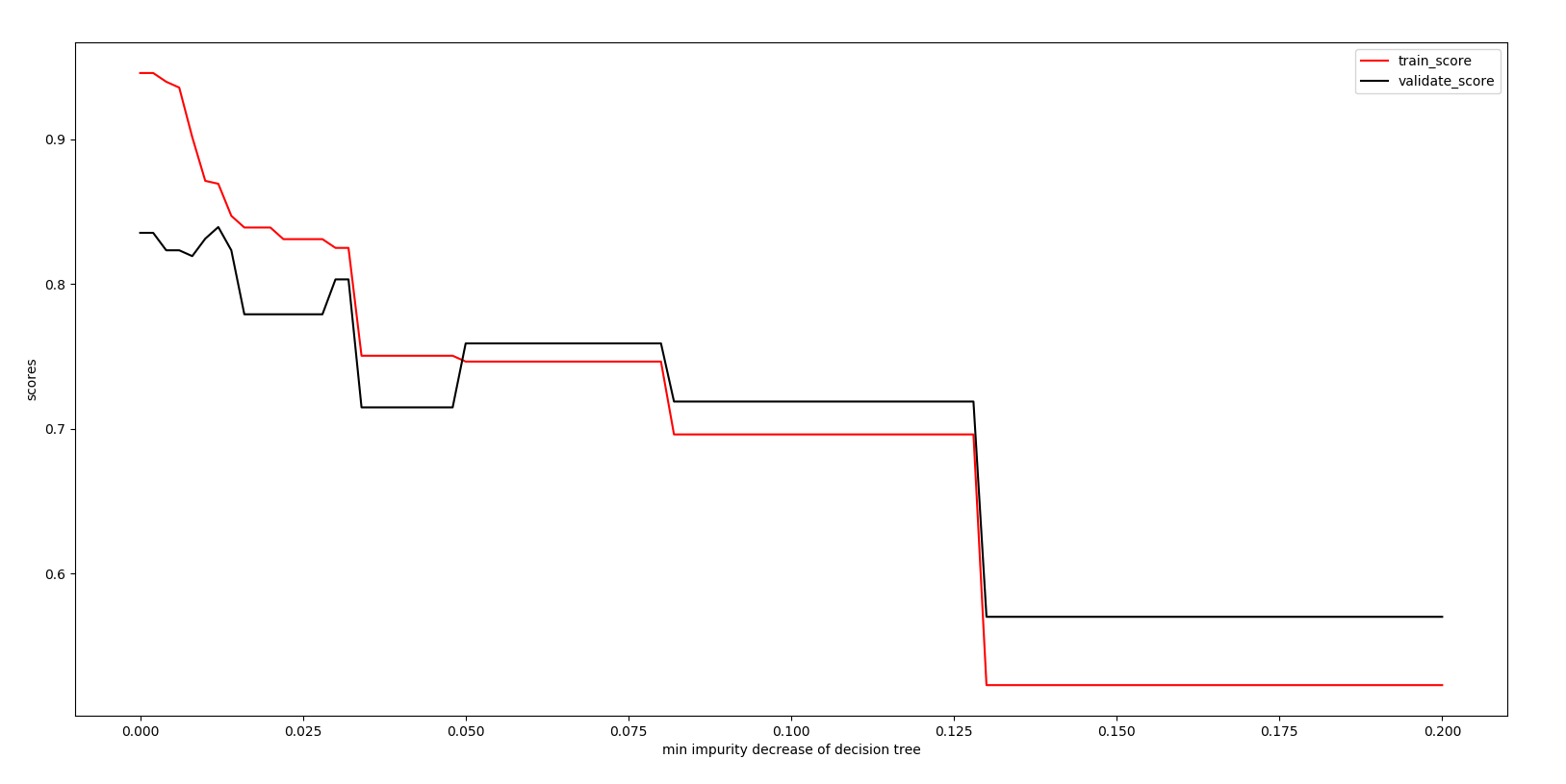
test score 3: 0.8215158924205379

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| entropy | 训练集 | 验证集 | 测试集 | | |
| 1625Data | 2/3的数据 | 1/3的数据 | 746Data | impensData | schillingData |
|  | 0.9457 | 0.8353 | 0.8425 | 0.7793 | 0.7940 |

#### 确定最优的min\_impurity\_decrease

数据集的纯度变化值满足一个阈值，才能进行分裂。（模型的最终学习目的，使得决策树的分支结点所包含的样本尽可能属于同一类别，即结点的纯度越来越高。

1）对\_746Data进行划分，选取2/3的数据集来学习模型， 选取1/3的数据集来作为验证集，其他数据集作为测试集。使用参数，criterion=‘entropy’， class\_weight=’balanced’,其他为默认参数，max\_path为上述实验结果，确定最优的min\_impurity\_decrease。



min\_impurity\_decrease: 0.0

train score: 0.9456740442655935

validate score: 0.8353413654618473

test score 1: 0.8424615384615385

test score 2: 0.7793030623020063

test score 3: 0.7940097799511002

----------------------------

min\_impurity\_decrease: 0.002

train score: 0.9456740442655935

validate score: 0.8353413654618473

test score 1: 0.8424615384615385

test score 2: 0.7793030623020063

test score 3: 0.7940097799511002

----------------------------

min\_impurity\_decrease: 0.004

train score: 0.9396378269617707

validate score: 0.8232931726907631

test score 1: 0.8498461538461538

test score 2: 0.7930306230200633

test score 3: 0.7863691931540342

----------------------------

min\_impurity\_decrease: 0.006

train score: 0.9356136820925554

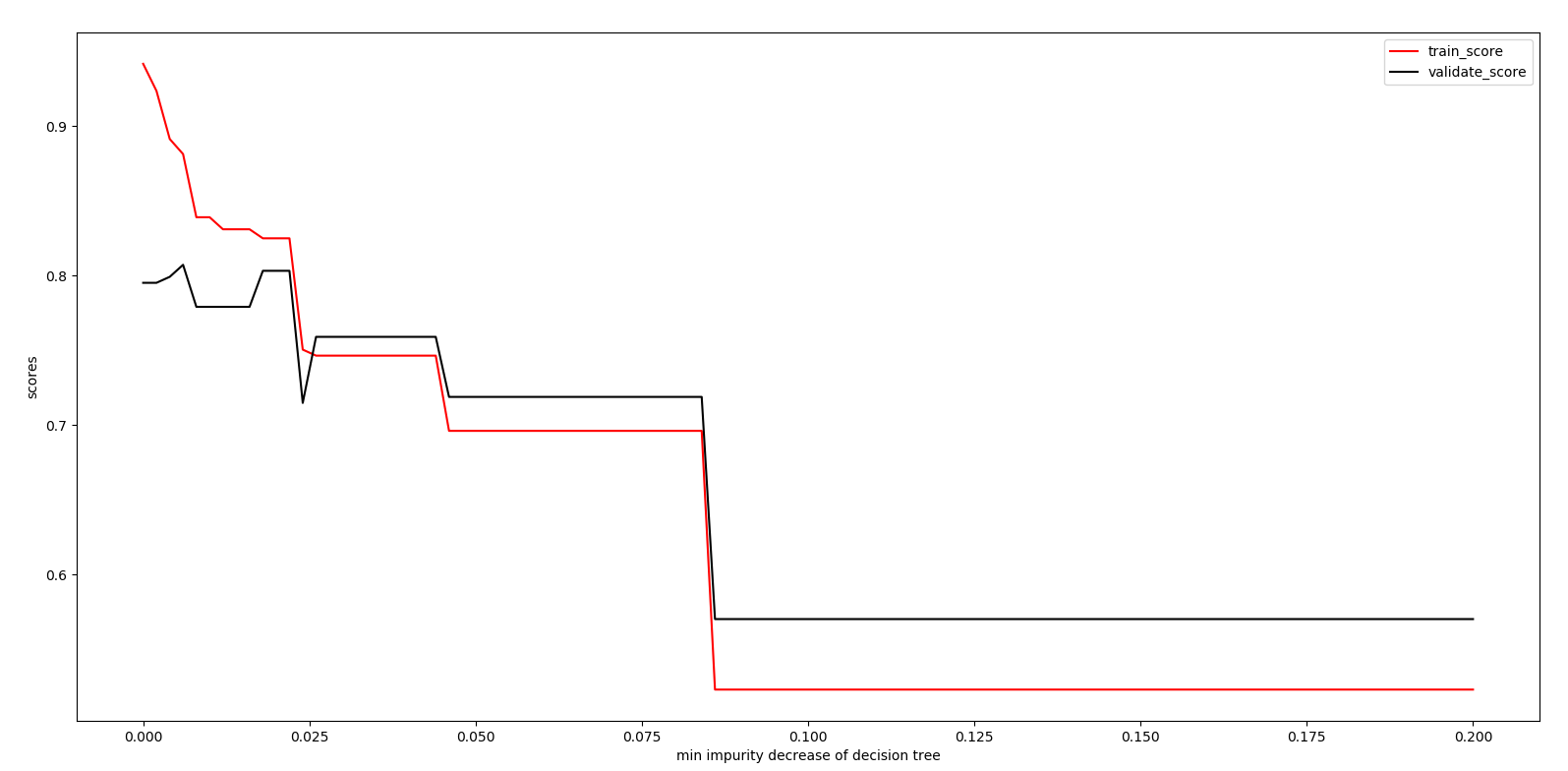
validate score: 0.8232931726907631

test score 1: 0.8498461538461538

test score 2: 0.7940865892291447

test score 3: 0.7863691931540342

2）对\_746Data进行划分，选取2/3的数据集来学习模型， 选取1/3的数据集来作为验证集，其他数据集作为测试集。使用参数，criterion=‘gini’， class\_weight=’balanced’,其他为默认参数，max\_path为上述实验结果，确定最优的min\_impurity\_decrease。



max\_depth: 8

train score: 0.9416498993963782

validate score: 0.7951807228915663

test score 1: 0.8732307692307693

test score 2: 0.7803590285110876

test score 3: 0.7930929095354523

min\_impurity\_decrease: 0.002

max\_depth: 8

train score: 0.9235412474849095

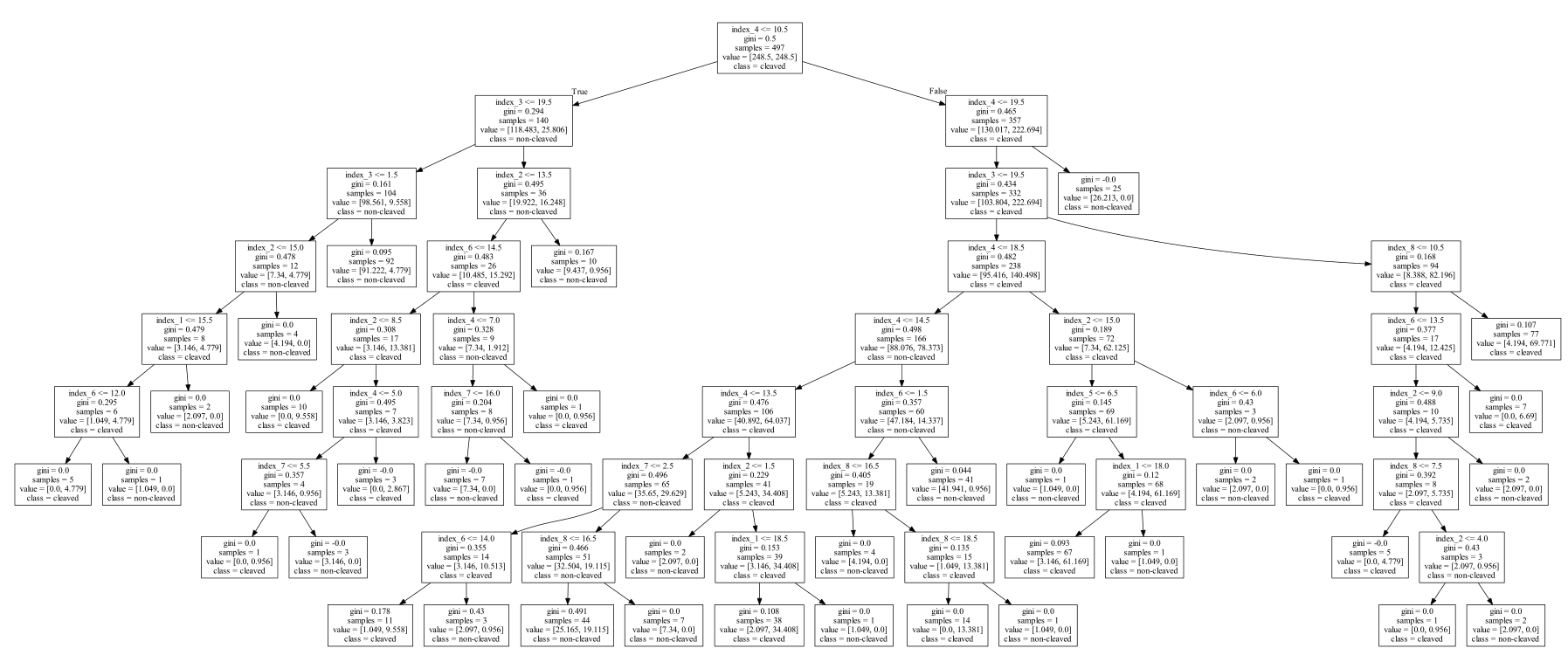
validate score: 0.7951807228915663

test score 1: 0.8904615384615384

test score 2: 0.8152059134107709

test score 3: 0.8215158924205379

根据上述实验，确定当以746Data为训练集时，



* 1. 利用GridSearchCV求最优参数



有进行了交叉验证，cv=5表示每次计算都把数据集分成5份，其中一份作为交叉验证数据集，其他作为训练集，从而得出最优参数。

### 2、手动实现

### 3、结论

相比于逻辑回归模型，决策树模型在面对不平衡类别的样本数据时，具有更好的泛化能力。

最优参数寻找过程，一般是确定一个参数后，找到另一个参数的最优解下的最优值。所以第一个参数的初始值设定会对另一个参数产生较大的影响